CHIMERYS: An AI-Driven Leap Forward in Peptide Identification

Martin Frejno[1]; Daniel P Zolg[1]; Tobias Schmidt[1]; Siegfried Gessulat[1]; Michael Graber[1]; Florian Seefried[1]; Magnus Rathke-Kuhnert[1]; Samia Ben Fredj[1]; Shyamnath Premnadh[1]; Patroklos Samaras[1], Kai Fritzemeier[2]; Frank Berg[2]; Waqas Nasir[2]; David Horn[3]; Bernhard Delanghe[2]; Christoph Henrich[2]; Bernhard Kuster[4]; Mathias Wilhelm[4] [1]MSAID GmbH, Garching b.München, Germany; [2]Thermo Fisher Scientific (Bremen) GmbH, Bremen, Germany; [3]Thermo Fisher Scientific, San Jose, CA; [4]Technical University of Munich, Freising, Germany

ABSTRACT

Purpose: Chimeric spectra represent a substantial challenge for bottom-up proteomics data analysis. Here, we describe CHIMERYS™, a novel, highly scalable, cloud-native, microservice-based and artificial intelligence-powered search algorithm that rethinks the analysis of tandem mass spectra from the ground up and deconvolutes chimeric spectra based on predicted fragment ion intensities.

Methods: We performed comparative analyses of standard HeLa tryptic digests that were acquired on various mass spectrometry platforms using different gradient lengths and isolation widths, as well as *in-silico* generated and publicly available datasets from various organisms using Sequest HT™, the Precursor Detector Node, INFERYS™ Rescoring [1] and CHIMERYS™ as implemented in a pre-release version of Thermo Scientific™ Proteome Discoverer™ 3.0 software.

Results: CHIMERYS doubles peptide identifications in classical data-dependent acquisition (DDA) datasets compared to Sequest HT and increases the number of identified peptides per protein by 2.5-fold on average, which translates to ~2 PSMs per spectrum and an identification rate of >80%. Entrapment analyses suggest that the CHIMERYS score set is well-calibrated and dilution experiments confirm that peptides unique to CHIMERYS follow the expected ratio distribution. Experiments based on simulated chimeric spectra establish that CHIMERYS has a sensitivity of >90%. Using CHIMERYS enables more efficient data acquisition strategies, as both wider isolation windows and shorter gradients can be used to generate more PSMs in a shorter timeframe.

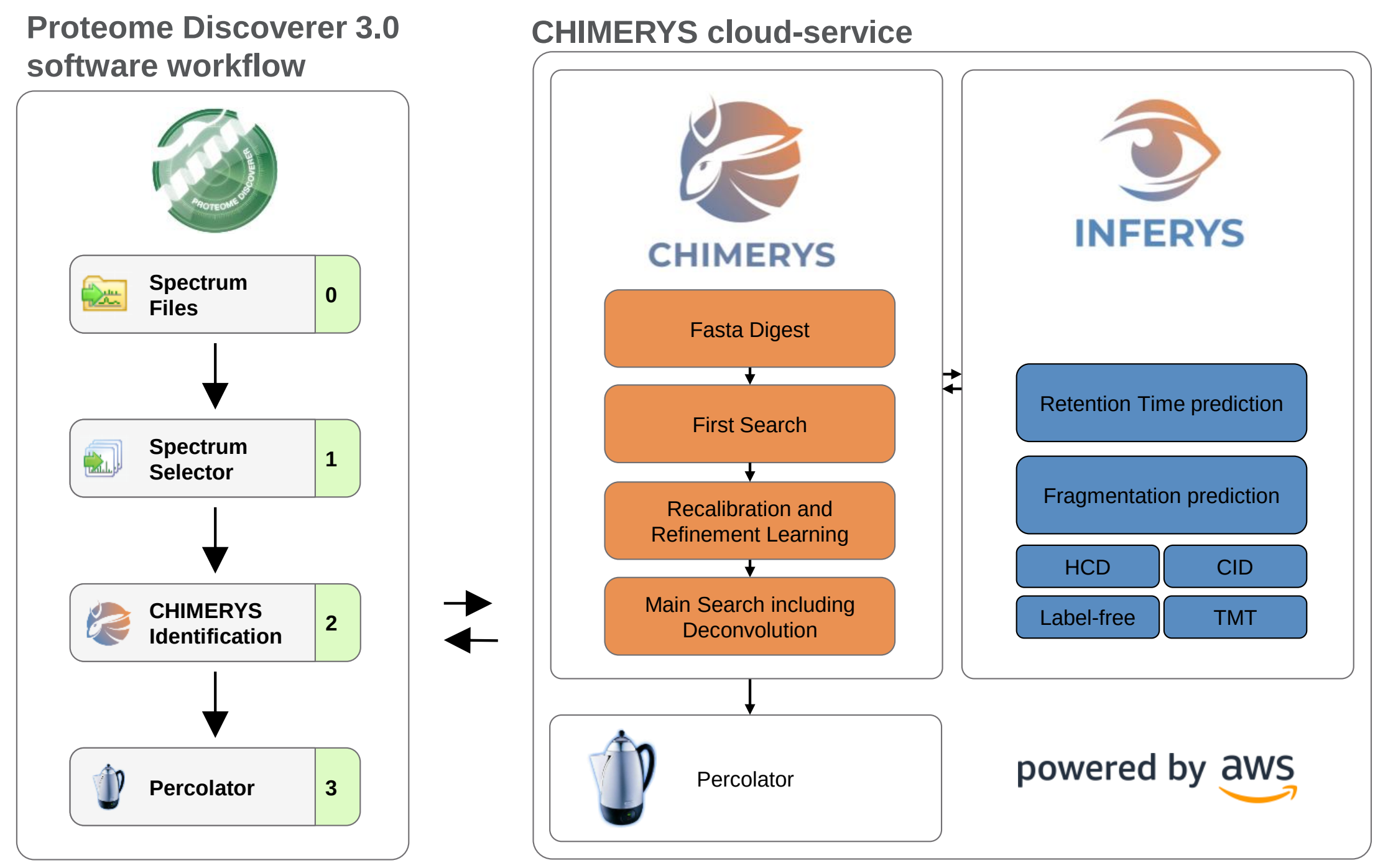
INTRODUCTION

Matching peptide sequences to tandem mass spectra is integral to bottom-up proteomics. Chimeric spectra are estimated to constitute >40% of DDA data [2], violating the assumption that one spectrum represents one peptide. Some search engines allow multi-pass searches or duplicate chimeric spectra for several possible precursors, but few account for the fact that the measured intensities of (isobaric) fragment ions may be the sum of multiple peptides. This introduces errors and leaves valuable information unused, resulting in far fewer peptide identifications than contained in the data. Here, we describe CHIMERYS, a new AI-based search algorithm that rethinks the analysis of tandem mass spectra from the ground up. It routinely doubles the number of peptide identifications in comparison to classical search algorithms and reaches identification rates of >80%.

MATERIALS AND METHODS

Data Analysis

CHIMERYS is a cloud-native search algorithm that uses accurate predictions of peptide fragment ion intensities and retention times provided by the deep learning framework INFERYS 2.0. Based on an initial coarse search, INFERYS performs data-driven model refinement to maximize prediction accuracy. Tandem mass spectra are analyzed without pre-processing or candidate selection using features detected in precursor mass spectra. Instead, all candidates in the isolation window of a given tandem mass spectrum are considered simultaneously and compete for measured fragment ion intensity in one concerted step. CHIMERYS aims to explain as much measured intensity with as few candidate peptides as possible, resulting in the deconvolution of chimeric spectra. Peptide-spectrum match (PSM)-level false discovery rate (FDR)-control is performed using Percolator [3]. CHIMERYS profits from cloud-based parallelization and is available through a node in a pre-release version Thermo Scientific Proteome Discoverer 3.0 software.



RESULTS

CHIMERYS doubles peptide identifications in single-shot full proteome DDA datasets

CHIMERYS' deconvolution algorithm identifies peptides hidden in chimeric spectra of DDA data files. Here, a digest of a HeLa cell lysate was analyzed using a 1-hour gradient on a Thermo Scientific™ Orbitrap Exploris™ 480 mass spectrometer and processed in Proteome Discoverer software using Sequest HT and CHIMERYS. The results demonstrate a more comprehensive data analysis when using CHIMERYS: over 80% of all MS2 spectra were matched to one or more peptide precursors and the average number of PSMs per spectrum substantially increases.

Figure 1. Number of PSMs, peptide and protein groups for a HeLa cell lysate

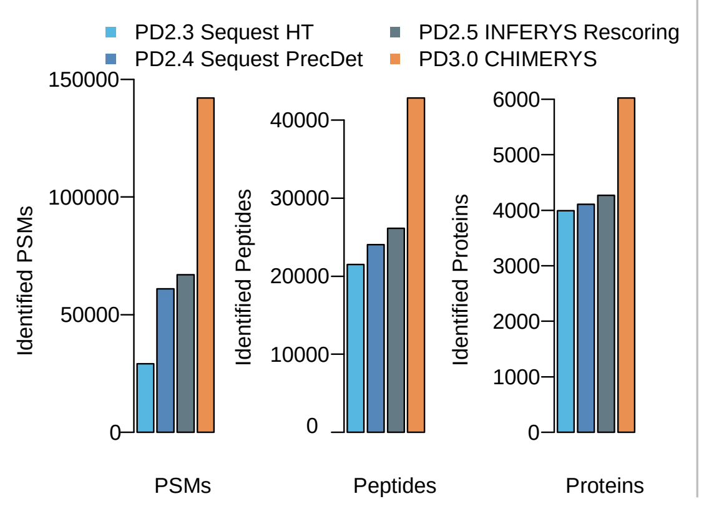


Figure 2. Overlap of peptide identifications between different search engines and processing workflows

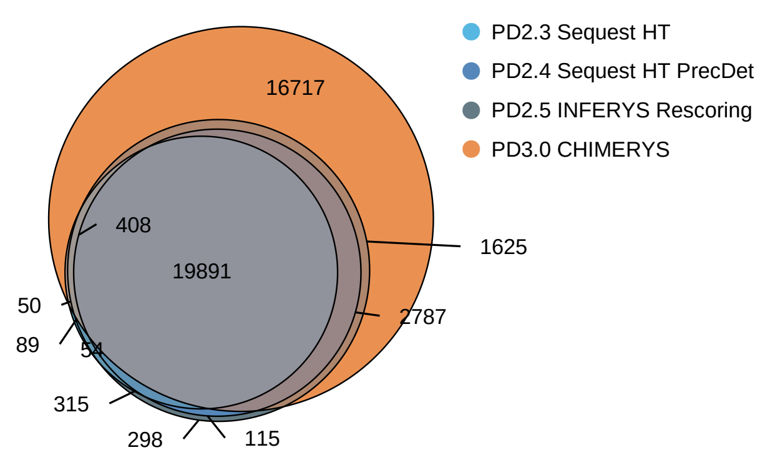


Figure 3. Number of PSMs per spectrum and identification rate achieved by CHIMERYS demonstrates the extent of the chimeric spectra problem in DDA data

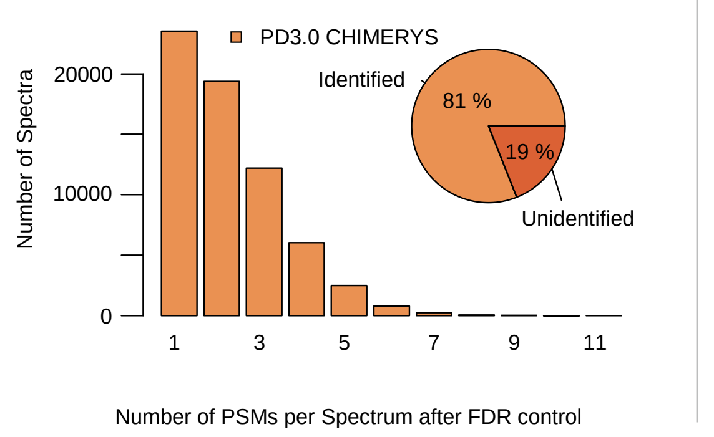
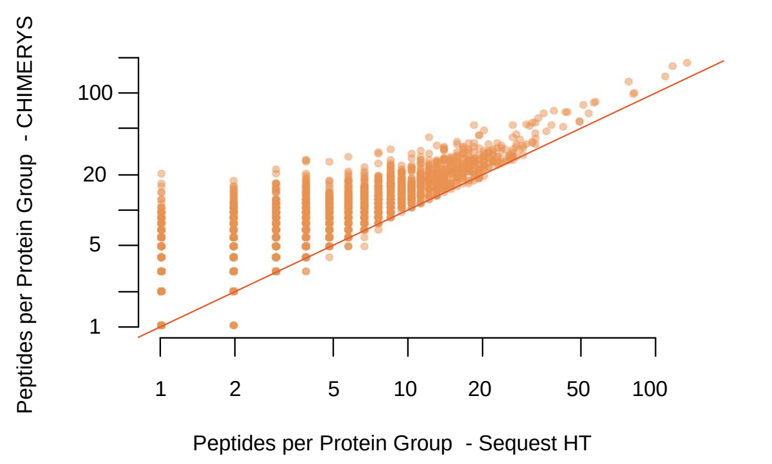


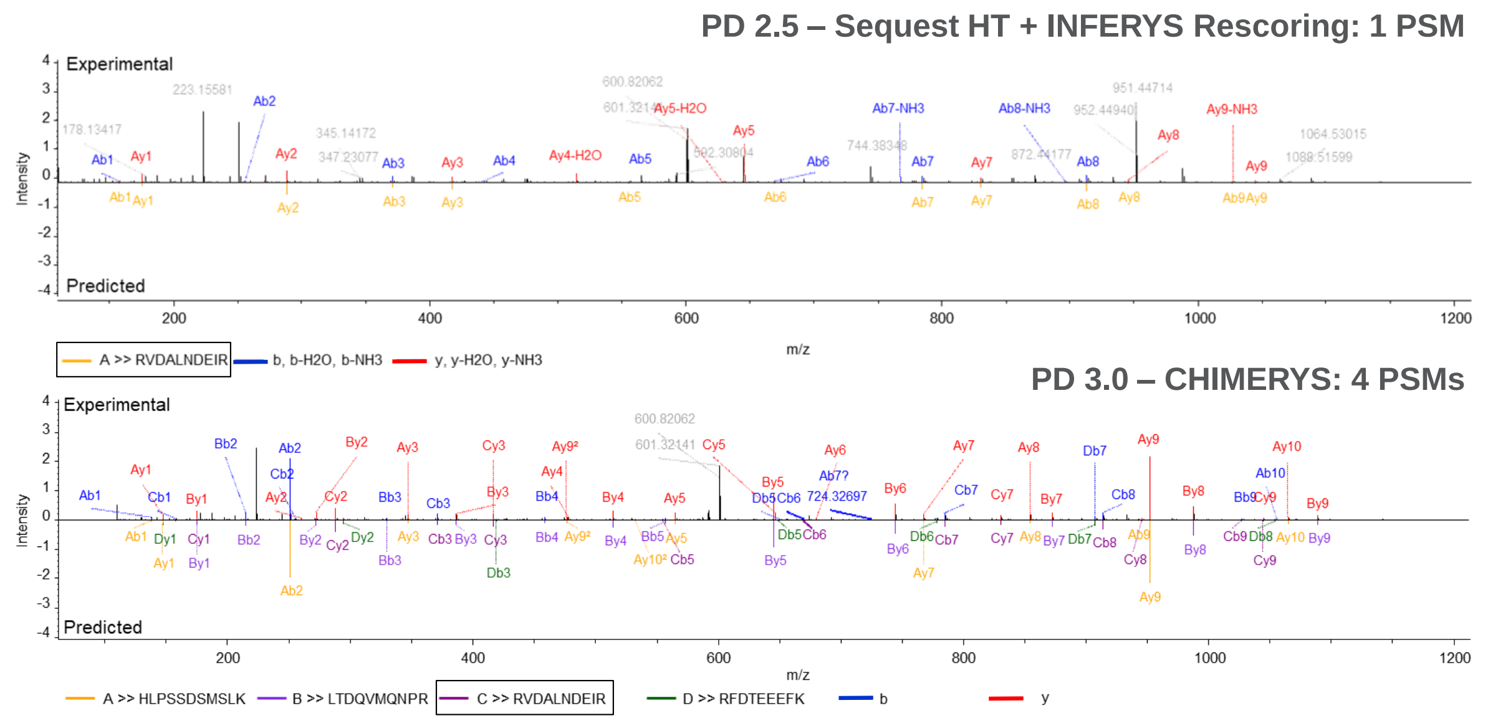
Figure 4. Number of peptides per protein group identified by CHIMERYS or Sequest HT demonstrates the increase in sequence coverage when using CHIMERYS



Accurate deconvolution by CHIMERYS unlocks information hidden in chimeric spectra

CHIMERYS deconvolutes MS2 spectra by considering all peptides for a given spectrum simultaneously, which then compete for the observed experimental intensity in a single step. This results in the identification of several PSMs from chimeric spectra. Using the Proteome Discoverer Spectrum Viewer functionality with direct connection to INFERYS 2.0, users can visualize the proportional contributions of the individual peptides for every single MS2 spectrum in a mirror plot.

Figure 5. Mirror plot of an experimental spectrum and PSMs identified by Sequest HT and INFERYS Rescoring (top panel) or CHIMERYS (bottom panel) at 1% FDR. While INFERYS Rescoring identifies only one peptide, CHIMERYS identifies three additional peptides, resulting in a drastically increased explained intensity of the experimental spectrum.



Validation of CHIMERYS' results using entrapment searches

Double-decoy approaches enable the calculation of an entrapment FDR and are common benchmarking methods to determine the correctness of FDR estimations. Here, we utilized a human database and appended 8 different plant databases (ratio ~1:7.5 proteins; shared peptides including I/L isomers were removed) to demonstrate the accuracy of the PSM-level FDR calculation performed by Percolator on CHIMERYS' search results of a 1h HeLa cell lysate measurement.

Figure 6. Analysis of entrapment FDR and calculated FDR using a ~8x non-homologous plant database across different search engines and workflows

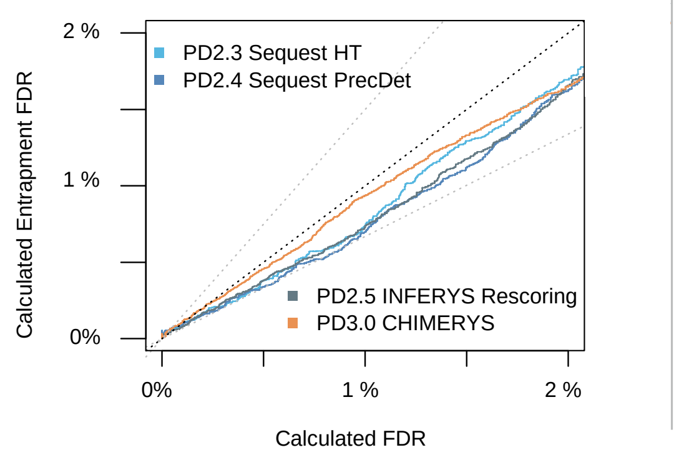
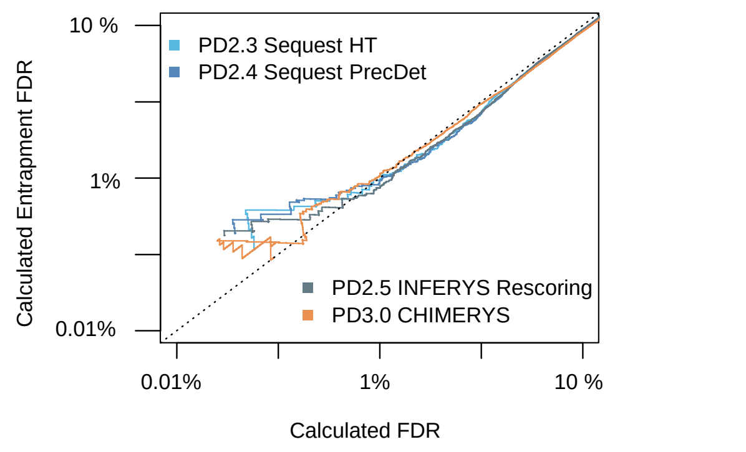


Figure 7. Double-log plot of the data shown in Figure 6. visualizing the low FDR region



CHIMERYS increases the number of accurately quantifiable peptides

Due to the increased analysis depth and comprehensive identification of PSMs and peptides, CHIMERYS aids in the accurate quantification of label-free datasets. We demonstrate this using a two-organism dilution series and compare the quantification results using the Minora feature detector node. This demonstrates that CHIMERYS produces more quantified peptides and proteins, especially low abundant ones. In this case, CHIMERYS quantifies 1.8-fold more proteins compared to Sequest HT.

Figure 8. Quantification of peptide ratios from a HeLa/Yeast dilution experiment (125ng/250ng)

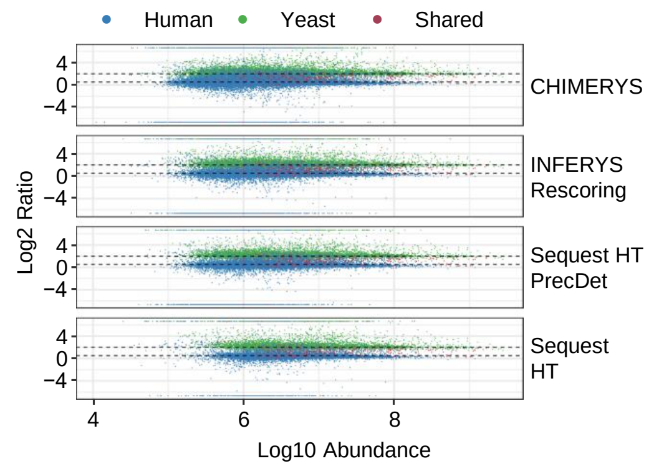
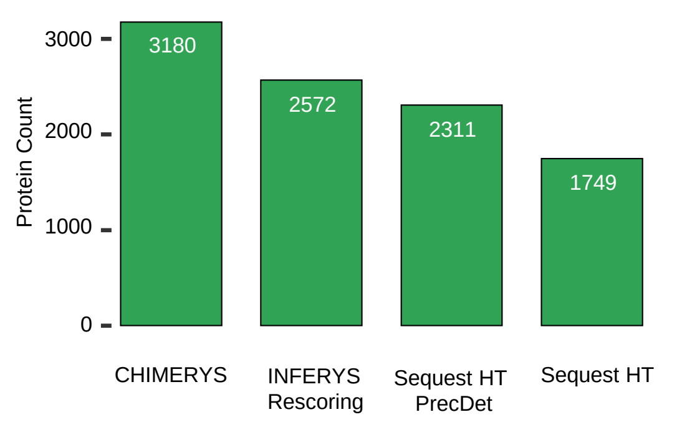


Figure 9. Number of quantified Yeast proteins from HeLa/Yeast dilution experiment



CHIMERYS demonstrates an exquisite sensitivity in simulation experiments

To validate CHIMERYS, we developed an *in-silico* chimeric spectra system (ICS) that spikes *in-silico* generated chimeric spectra into raw files, which can then be used as a ground truth dataset to evaluate search algorithms. Briefly, the system selects seed MS2 spectra with high-confident identifications from a prior database search from a raw file and convolutes them with several predicted MS2 spectra. To create realistic chimeric data, predicted spectra are derived from peptides with a precursor m/z value within the isolation window of the seed MS2 spectrum and a similar predicted retention time. The created raw file is then submitted to both CHIMERYS and Sequest HT. Using this system, we demonstrate the sensitivity of CHIMERYS, which recovers >91% of the *in-silico* chimeric spectra in the convoluted data.

Figure 10. Schema of the ICS system for generating a ground-truth dataset containing *in-silico* chimeric spectra

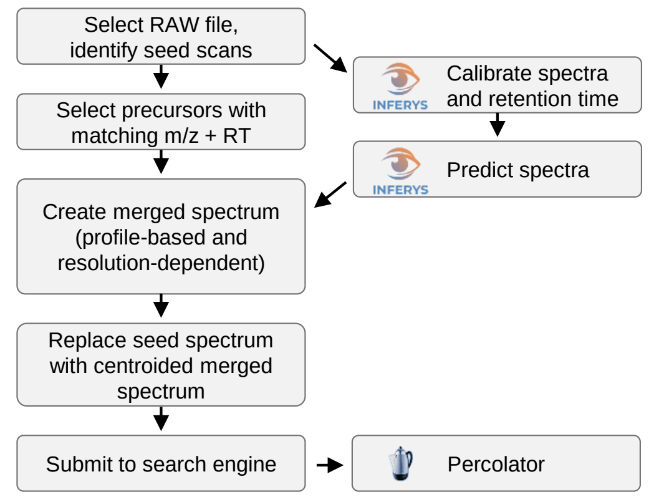
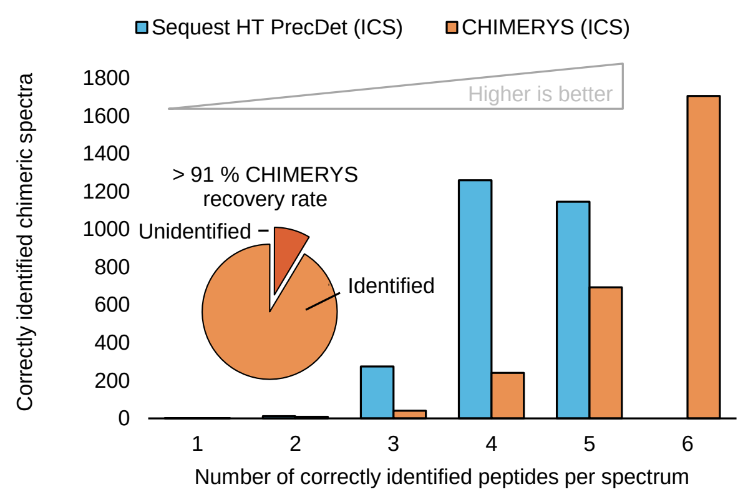

Select RAW file, identify seed scans
Select precursors with matching m/z + RT
Create merged spectrum (profile-based and resolution-dependent)
Replace seed spectrum with centroided merged spectrum
Submit to search engine
INFERYS
Calibrate spectra and retention time
Predict spectra
Percolator

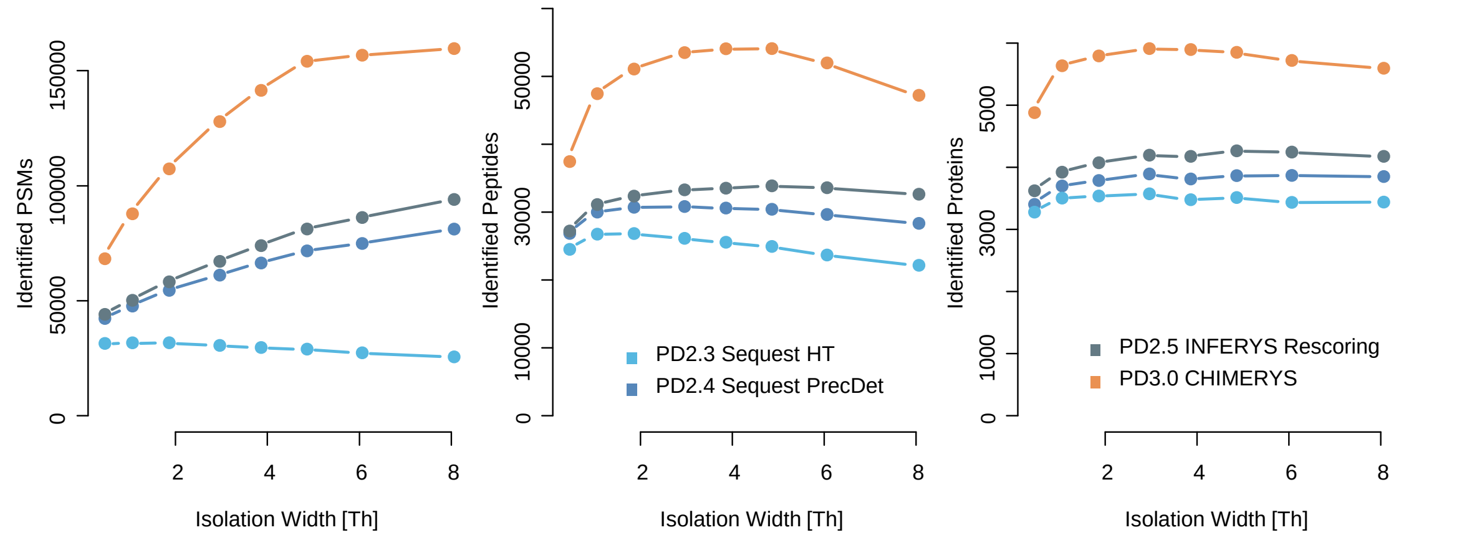
Figure 11. Recovery of 2,700 *in-silico* generated chimeric spectra (6 peptides each) by CHIMERYS and Sequest HT at 1% FDR level



CHIMERYS enables optimized acquisition settings and profits from increased MS2 complexity

CHIMERYS' deconvolution algorithm is compatible with highly complex samples resulting in convoluted MS2 spectra. Hence, it allows for optimizing data acquisition settings to increase measurement efficiency by identifying more proteins per unit time. Here, we demonstrate that CHIMERYS enables wider DDA isolation windows that result in more chimeric MS2 spectra, providing more identifications and better sequence coverage.

Figure 12. Number of PSMs, peptide and protein groups identified from a DDA HeLa cell lysate digest acquired on a Thermo Scientific™ Orbitrap Eclipse™ Tribrid™ mass spectrometer using a 1-hour gradient and MS2 isolation windows between 0.4 Th and 8 Th.



CHIMERYS enables increasing sample throughput by using shorter chromatography

CHIMERYS uniquely deciphers complex samples and MS2 spectra, enabling shorter gradients for LC-MS/MS measurements without losing peptide or protein information in comparison to Sequest HT. Here, we demonstrate how CHIMERYS identifies the same number of peptides and protein groups in 1/3 of the measurement time.

Figure 13. Number of PSMs, peptide and protein groups identified by CHIMERYS or Sequest HT from digests of a HeLa cell lysate acquired on an Orbitrap Exploris 480 MS with gradient lengths ranging from 8 to 60 minutes on a Thermo Scientific™ Vanquish™ Neo UHPLC system

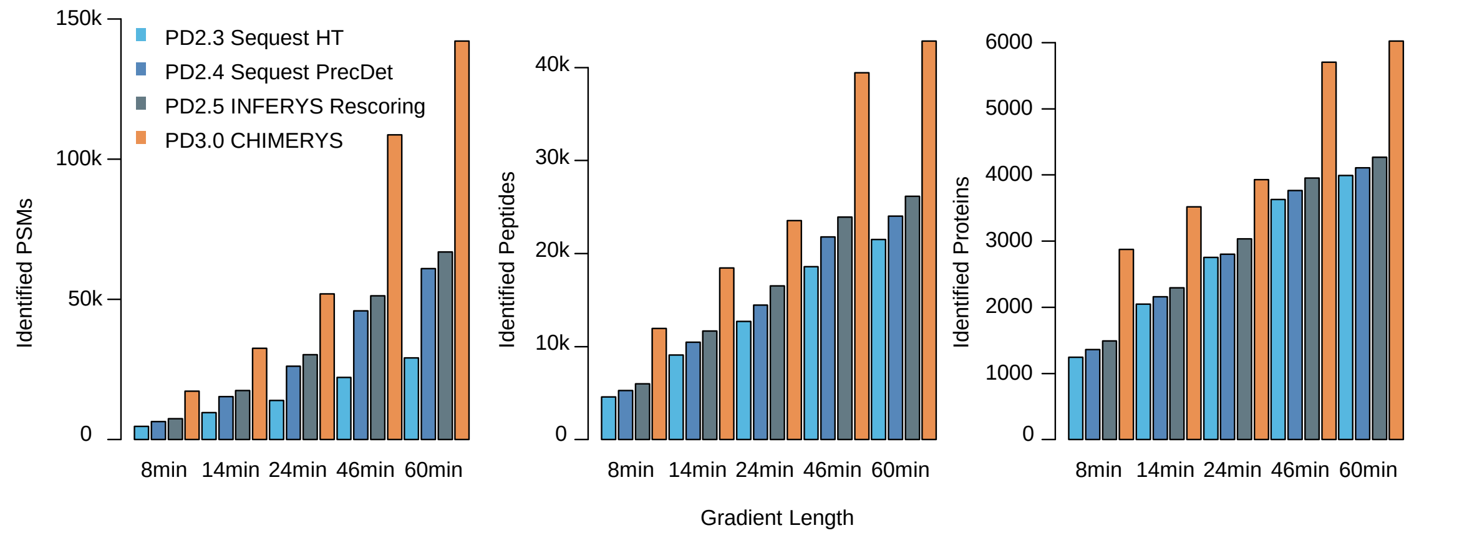
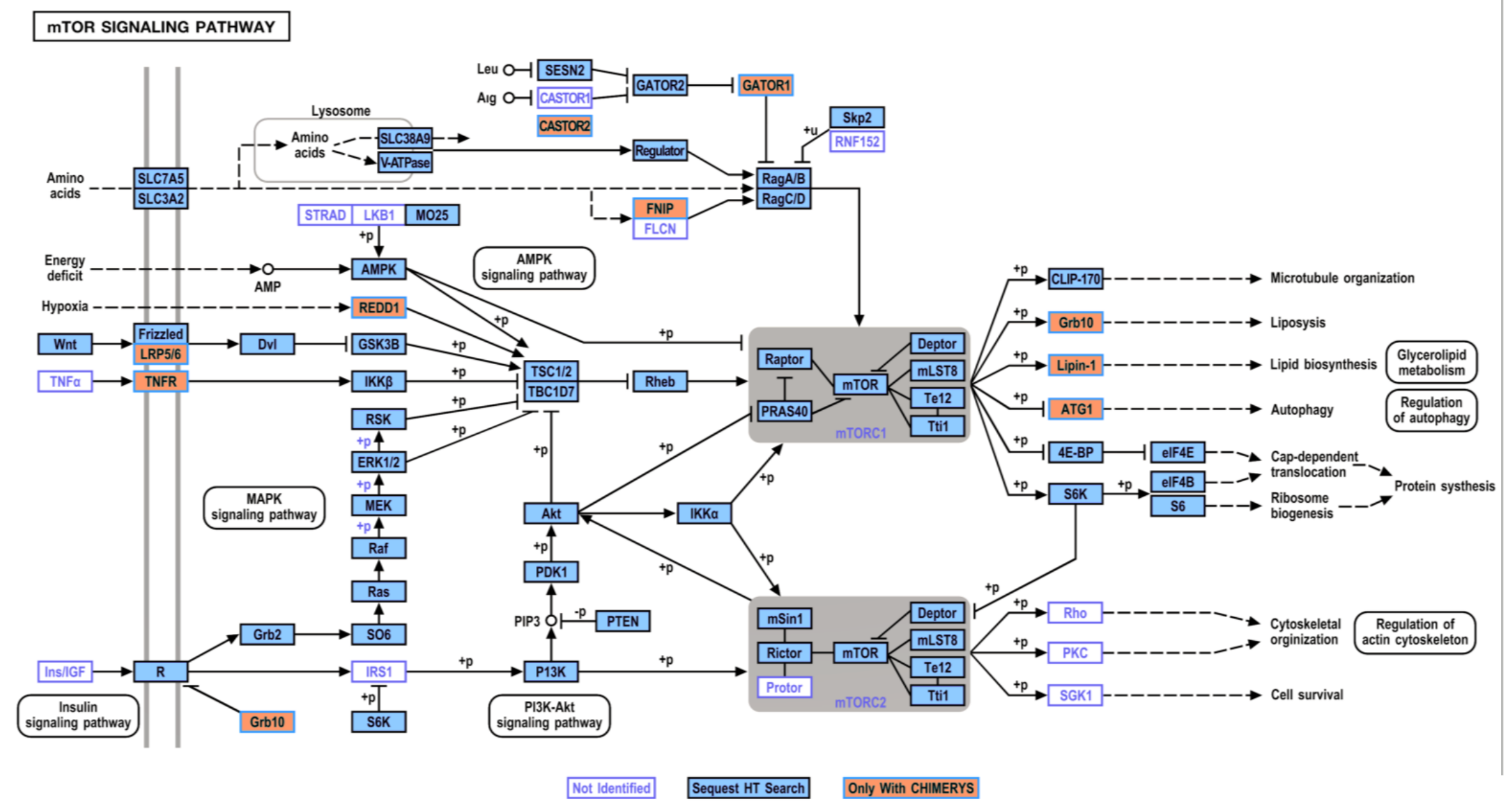


Figure 14. Proteins of the mTOR signaling pathway identified by CHIMERYS or Sequest HT in a HeLa cell lysate demonstrate the potential for new biological insight through extended protein and pathway coverage that can be generated using CHIMERYS



CHIMERYS outperforms Sequest HT on datasets from different biological sources

CHIMERYS is fueled by predictions from INFERYS 2.0 that are independent of the sample source under investigation. Paired with its resilience with respect to highly complex data, CHIMERYS is well-equipped to handle fractionated or non-fractioned measurements from organisms from all kingdoms of life [4] and less complex samples like body fluids [5]. Here, we demonstrate its capabilities on a selection of publicly available data.

Figure 15. Protein groups identified by CHIMERYS and Sequest HT for a fractionated *Arabidopsis thaliana* proteome; raw data from PRIDE Project PXD019483 [4]

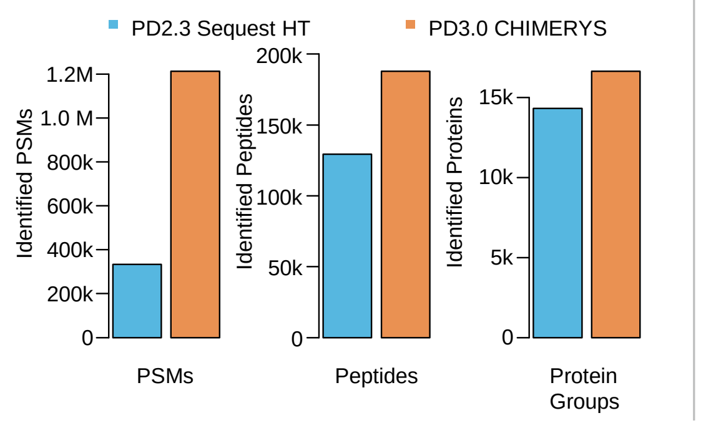
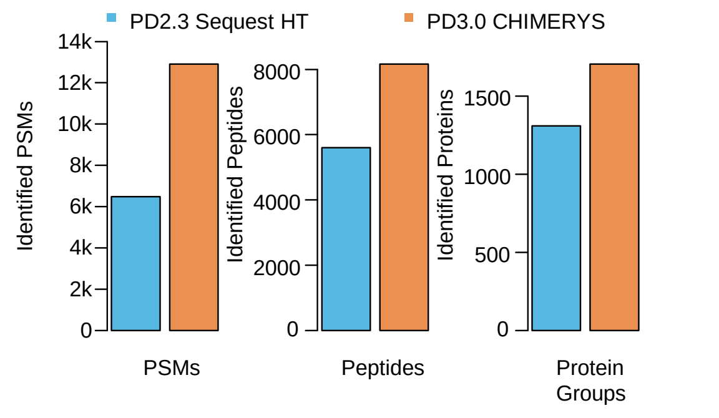


Figure 16. Protein groups identified by CHIMERYS and Sequest HT for a single 30 min Urine proteome file; raw data from PRIDE Project PXD015087 [5]



CONCLUSIONS

- CHIMERYS is an innovative, cloud-native search algorithm that uses AI-based predictions to deconvolute chimeric spectra and is fully integrated into Proteome Discoverer 3.0 software
- Using CHIMERYS results in drastically increased numbers of PSM, peptide and protein group identifications, higher sequence coverage and more confident quantification
- CHIMERYS excels at analyzing complex samples, enabling more efficient measurements, advanced acquisition settings and shorter gradients to enhance proteomic throughput, productivity and efficiency

REFERENCES

1. Zolg, DP; Gessulat, S; Paschke, C, Frejno M, et al. INFERYS rescoring: Boosting peptide identifications and scoring confidence of database search results. *Rapid Commun Mass Spectrom*. 2021;e9128. https://doi.org/10.1002/rcm.9128
2. Dorfer V; Maltsev S; Winkler S; Metchler K. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *Journal of Proteome Research* 2018 17 (8), 2581-2589. https://doi.org/10.1021/acs.jproteome.7b00836
3. The M; MacCoss MJ; Noble WS; Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J Am Soc Mass Spectrom.* 2016;27(11):1719-1727. https://doi.org10.1007/s13361-016-1460-7
4. Müller JB; Geyer, PE; Colaço, A.R, Mann M; et al. The proteome landscape of the kingdoms of life. *Nature* 582, 592–596 (2020). https://doi.org/10.1038/s41586-020-2402-x
5. Bian, Y; Zheng, R; Bayer, FP; Kuster B; et al. Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC–MS/MS. *Nat Commun* 11, 157 (2020). https://doi.org/10.1038/s41467-019-13973-x

ACKNOWLEDGEMENTS

The authors would like to thank the alpha and beta testers and especially Karl Mechtler for the evaluation and validation of CHIMERYS. The authors wish to thank numerous scientists and colleagues at MSAID and Thermo Fisher Scientific for fruitful discussions and technical assistance.

TRADEMARKS/LICENSING

PO66098